



DTIC FILE COPY

TEXAS A&M UNIVERSITY  
COLLEGE STATION, TEXAS 77843-3143

2

Department of STATISTICS  
Statistical Interdisciplinary  
Research Laboratory  
F415EP@TAMU.EDU BITNET

Emanuel Parzen  
Distinguished Professor  
Phone 409-845-3188  
Fax 409-845-3144

AD-A224 317

TIME SERIES, STATISTICS, AND INFORMATION

Emanuel Parzen  
Department of Statistics  
Texas A&M University

Technical Report No. #113  
June 1990

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input checked="" type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	



Texas A&M Research Foundation  
Project No. 6547

'Functional Statistical Data Analysis and Modeling'

Sponsored by the U. S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited

DTIC  
ELECTE  
JUL 20 1990  
S E D

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER  AR0 27574.3 .nA	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle)  Time Series, Statistics, and Information		5. TYPE OF REPORT & PERIOD COVERED  Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Emanuel Parzen		8. CONTRACT OR GRANT NUMBER(s)  DAAL03-90-G-0069
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Institute of Statistics College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS  •		12. REPORT DATE June 1990
		13. NUMBER OF PAGES 34
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report)  Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)  NA		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Time series analysis, Statistical culture, entropy, cross-entropy, Renyi information, Comparison density functions, Approximation of positive functions, Equivalence and orthogonality of normal time series, estimation of spectral densities, Tail classification, Probability laws, Spectral densities, Exploratory analysis of time series.		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) This paper is a broad survey of ideas for the future development of statistical methods of time series analysis based on investigating the many levels of relationships between time series analysis, statistical methods unification, and inverse problems with positivity constraints. It is hoped that developing these relations will: help integrate old and new directions of research in time series analysis; provide research tools for applied and theoretical statisticians in the 1990's and coming era of statistical information; make possible unification of statistical methods and the development of Statistical Culture. New results include a new information divergence between spectral density functions. To be presented at Workshop in New Directions in Time Series Analysis at Institute for Mathematics and its Applications, University of Minnesota on July 16, 1990.		

DD FORM 1 JAN 73 1473

EDITION OF 1 NOV 68 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

# TIME SERIES, STATISTICS, AND INFORMATION

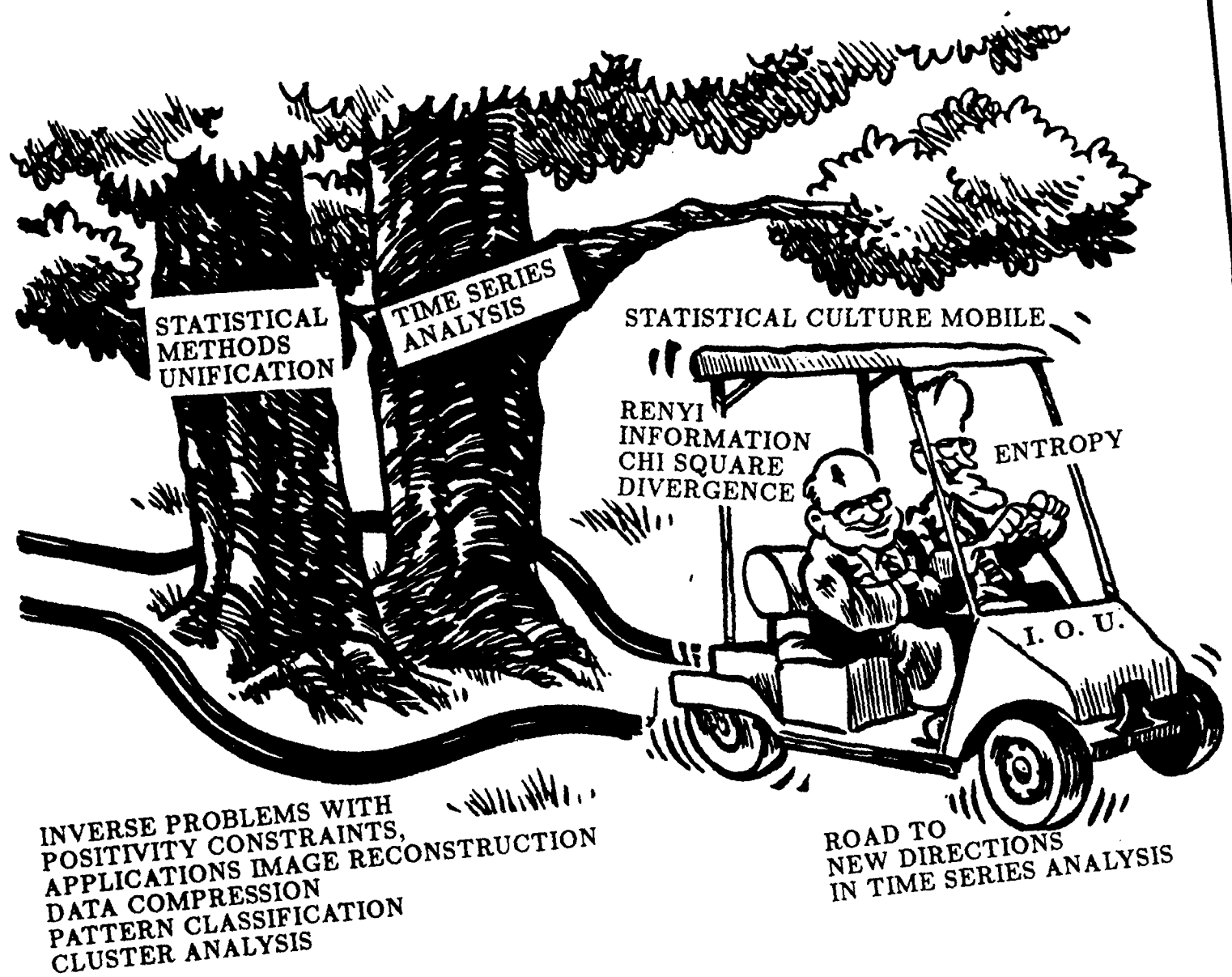
by

Emanuel Parzen

Department of Statistics, Texas A&M University

Abstract by Image

## STATISTICAL CULTURE BY I.O.U. (INFORMATION OPTIMIZATION UNIFICATION)



# TIME SERIES, STATISTICS, AND INFORMATION

by

Emanuel Parzen<sup>1</sup>

Department of Statistics, Texas A&M University

## Abstract

This paper is a broad survey of ideas for the future development of statistical methods of time series analysis based on investigating the many levels of relationships between time series analysis, statistical methods unification, and inverse problems with positivity constraints. It is hoped that developing these relations will: help integrate old and new directions of research in time series analysis; provide research tools for applied and theoretical statisticians in the 1990's and coming era of statistical information; make possible unification of statistical methods and the development of Statistical Culture. New results include a new information divergence between spectral density functions. Topics discussed include:

- 1) Traditional entropy and cross-entropy,
- 2) Renyi and Chi-square information divergence,
- 3) Comparison density functions,
- 4) Approximation of positive functions (density functions) by minimum information divergence (maximum entropy),
- 5) Equivalence and Orthogonality of Normal Time Series,
- 6) Asymptotic Information of Stationary Normal Time Series,
- 7) Estimation of Finite Parameter Spectral Densities,
- 8) Minimum information estimation of spectral densities and power index correlations,
- 9) Tail classification of probability laws and spectral densities, AND
- 10) Sample Brownian Bridge exploratory analysis of time series. (KF) ←

---

<sup>1</sup>Preliminary version of paper to be presented at Workshop in New Directions in Time Series Analysis at Institute for Mathematics and Its Applications, University of Minnesota on July 16, 1990, a day dedicated to honor John Tukey's contributions to Time Series Analysis. Research supported by U. S. Army Research Office.

## 0. Introduction

The general level of the current relation (or non-relation) between statistics and time series analysis is: (1) many applied statisticians are ignorant about the theory of time series analysis, (2) many departments of statistics offer almost no courses in time series analysis (often relying on courses taught in economics or engineering departments), (3) theoretical statisticians traditionally have regarded time series analysis as safe to ignore because it is a "technical" subject in which it is difficult to confront basic issues of statistical inference which are the problems about which they want to do research.

Time series models are becoming of research interest to some theoretical statisticians whose primary research areas involve statistical analysis of data obeying the classical model of independent observations. They would like to investigate the extension of their work to data obeying probability models of dependence. This relation of time series analysis and statistics has a possibility of being superficial because it is completely methods-driven, rather than problem-driven. Consequently it may not handle problems that are of real interest to applied users of time series analysis. Questions of asymptotic rate of convergence of parameter estimators are technical problems which fail to treat basic problems (such as model identification and/or non-regular estimation for long-memory or non-Gaussian time series) that are usually the central problems in a time series analysis.

I propose that the narrow reasons why statisticians should learn about the methods of time series analysis (they are important for applications and many potential clients have time series problems) should be supplemented by broad information age reasons; the development of Statistical Culture requires that statisticians should learn about the theory of time series because it will help them improve their mastery of the basic methods of statistical analysis for traditional data consisting of independent observations. The theory of time series analysis needs to become *exoteric* (belonging to the outer or less initiate circle) as well as *esoteric*.

My concept of Statistical Culture (Parzen (1990)) proposes that a statistical analysis

should aim to provide not a single answer but a choice of answers (answers by several methods for the same problem). Therefore a framework for comparing answers is required. A framework should also provide ways of thinking for classical problems that extends to as many modern problems as possible. This paper discusses: the basic ideas of a unification of statistical methods in terms of information concepts, the relation between time series and statistics in terms of their relations with information statistics, a framework for time series methods in terms of information concepts, and suggestions for research problems in time series analysis. The references aim to include many influential papers on information methods in statistics; additional references are warmly solicited.

An aim of this paper is to stimulate discussion of the mind-boggling discovery which appears to be emerging in modern statistical research and which I call I. O. U. (Information, Optimization, Unification). It appears that one can find a common type of optimization and approximation problem that provides a link among almost all classical and modern statistical analysis problems! An inner product  $\langle f, g \rangle$  is the integral of the product  $fg$ . Let the information about an unknown NON-NEGATIVE function  $f$  be the values of linear functionals which are inner products of  $f$  with specified score functions  $J_k$ : for  $k = 1, \dots, m$ ,

$$\langle f, J_k \rangle = \tau_k$$

for specified constants  $\tau_k$  called "moment parameters". Find a NON-NEGATIVE function, denoted

$$\hat{f} \text{ or } \hat{f}(\tau_1, \dots, \tau_m),$$

which among all functions satisfying the above constraints minimizes an information divergence criterion (which includes as a special case maximum entropy). This problem is called an inverse problem with positivity constraints.

We favor Renyi information criteria which imply that  $\hat{f}$  has a representation  $f_\theta$ : for

suitable index  $\lambda$  and parameters  $\theta_k$  (which we call inverse parameters)

$$(f_\theta^\lambda - 1) / \lambda = \sum_{k=1}^m \theta_k J_k$$

The inverse parameters  $\theta$  are functions of the moment parameters and obtained by solving  $\langle f_\theta, J_k \rangle = \tau_k$ .

Uncertainty (probability and statistics) enters the picture because one observes a raw random function (denoted  $f^\sim$ ) or at least its inner products

$$\langle f^\sim, J_k \rangle = \tau_k^\sim, k = 1, \dots, m.$$

One then seeks  $f^\sim$  which is non-negative and minimizes the specified information criterion among all functions satisfying  $\langle f, J_k \rangle = \tau_k^\sim, k = 1, \dots, m$ . Among new data analytic tools that are open problems for research are "profile functions", defined as the minimum value of the criterion as a function of the moment parameters.

The problems that need to be solved to apply the foregoing approach to unifying statistical methods include

(1) introducing suitable density function  $d(u)$ ,  $0 < u < 1$ , whose estimation underlies conventional problems,

(2) determining sufficient statistics  $J_k(u)$  whose inner products with  $d(u)$  or  $d^\lambda(u)$  for some power  $\lambda$ , are regarded as most significantly different from zero and therefore provide the constraints on the unknown  $d$ ,

(3) determining information measures whose index  $\lambda$  provides a parameter formula for  $d$  of the form  $d^\lambda(u)$  is a linear combination of known functions  $J_k(u)$  with coefficients  $\theta_k$  to be estimated,

(4) developing and implementing algorithms to compute the solutions of the optimization problems.

Other aspects of unification of time series analysis methods were discussed in Parzen (1958), Parzen (1961), Parzen (1965), Parzen (1971), and Parzen (1974).

## 1. Traditional Entropy and Cross-Entropy

The (Kullback-Liebler) information divergence between two probability distributions  $F$  and  $G$  is defined (Kullback (1959)) by (our definitions differ from usual definitions by a factor of 2)

$$I(F; G) = (-2) \int_{-\infty}^{\infty} \log\{g(x)/f(x)\} f(x) dx$$

when  $F$  and  $G$  are continuous with probability density functions  $f(x)$  and  $g(x)$ ; when  $F$  and  $G$  are discrete, with probability mass functions  $p_F(x)$  and  $p_G(x)$ , information divergence is defined by

$$I(F; G) = (-2) \sum \log\{p_G(x)/p_F(x)\} p_F(x).$$

An information decomposition of information divergence is

$$I(F; G) = H(F; G) - H(F),$$

in terms of entropy  $H(F)$  and cross-entropy  $H(F; G)$ :

$$\begin{aligned} H(F) &= (-2) \int_{-\infty}^{\infty} \{\log f(x)\} f(x) dx, \\ H(F; G) &= (-2) \int_{-\infty}^{\infty} \{\log g(x)\} f(x) dx. \end{aligned}$$

## 2. Renyi and Chi Square Information

Adapting the fundamental work of Renyi (1961) this section offers a new definition of



Renyi information of index  $\lambda$ . For continuous  $F$  and  $G$ : for  $\lambda \neq 0, -1$

$$\begin{aligned}
 IR_{\lambda}(F; G) &= \frac{2}{\lambda(1+\lambda)} \log \int \left\{ \frac{g(y)}{f(y)} \right\}^{1+\lambda} f(y) dy \\
 &= \frac{2}{\lambda(1+\lambda)} \log \int \left\{ \frac{g(y)}{f(y)} \right\}^{\lambda} g(y) dy \\
 IR_0(F; G) &= 2 \int \left\{ \frac{g(y)}{f(y)} \log \frac{g(y)}{f(y)} \right\} f(y) dy \\
 &= 2 \int \left\{ \frac{g(y)}{f(y)} \log \frac{g(y)}{f(y)} - \frac{g(y)}{f(y)} + 1 \right\} f(y) dy \\
 IR_{-1}(F; G) &= -2 \int \left\{ \log \frac{g(y)}{f(y)} \right\} f(y) dy \\
 &= -2 \int \left\{ \log \frac{g(y)}{f(y)} - \frac{g(y)}{f(y)} + 1 \right\} f(y) dy
 \end{aligned}$$

An analogous definition holds for discrete  $F$  and  $G$ .

The second definition provides: (1) extensions to non-negative functions which are not densities, and also (2) a non-negative integrand which can provide diagnostic measures at each value of  $y$ .

Renyi information, for  $-1 < \lambda < 0$ , is equivalent to Bhattacharyya distance (Bhattacharyya (1943)).

In addition to Renyi information divergence (an extension of information statistics) one uses as information divergence between two non-negative functions an extension of chi-square statistics which has been developed by Read and Cressie (1988). For  $\lambda \neq 0, -1$ , Chi-square divergence of index  $\lambda$  is defined for continuous  $F$  and  $G$  by

$$C_{\lambda}(F; G) = \int B_{\lambda} \left( \frac{g(y)}{f(y)} \right) f(y) dy$$

where

$$\begin{aligned}
 B_{\lambda}(d) &= \frac{2}{(1+\lambda)} \left\{ d \left( \frac{d^{\lambda} - 1}{\lambda} \right) - d + 1 \right\} \\
 B_0(d) &= 2 \{ d \log d - d + 1 \} \\
 B_{-1}(d) &= -2 \{ \log d - d + 1 \}
 \end{aligned}$$

Important properties of  $B_\lambda(d)$  are:

$$B_\lambda(d) \geq 0, B_\lambda(1) = B'_\lambda(1) = 0,$$

$$B'_\lambda(d) = \frac{2}{\lambda} (d^\lambda - 1), B''_\lambda(d) = 2d^{\lambda-1}$$

$$B_1(d) = (d - 1)^2$$

$$B_0(d) = 2(d \log d - d + 1)$$

$$B_{-.5}(d) = 4(d^{.5} - 1)^2$$

$$B_{-1}(d) = -2(\log d - d + 1)$$

$$B_{-2}(d) = d(d^{-1} - 1)^2$$

An analogous definition holds for discrete  $F$  and  $G$ . Axiomatic derivations of information measures similar to  $C_\lambda$  are given by Jones and Byrne (1990).

The Renyi information and chi-square divergence measures are related:

$$IR_0(F; G) = C_0(F; G)$$

$$IR_{-1}(F; G) = C_{-1}(F; G)$$

For  $\lambda \neq 0, -1$ ,

$$IR_\lambda(F; G) = \frac{2}{\lambda(1 + \lambda)} \log \left\{ 1 + \left( \frac{\lambda(1 + \lambda)}{2} \right) C_\lambda(F; G) \right\}$$

Interchange of  $F$  and  $G$  is provided by the Lemma:

$$C_\lambda(F; G) = C_{-(1+\lambda)}(G; F)$$

$$IR_\lambda(F; G) = IR_{-(1+\lambda)}(G; F)$$

Our survey in this paper suggests (in section 6) a new class of information measures:

$$A_\lambda(F; G) = \int A_\lambda \left( \frac{g(y)}{f(y)} \right) f(y) dy$$

For  $\lambda \neq 0, -1$ , perhaps most usefully  $-1 < \lambda < 0$ ,

$$A_\lambda(d) = \{2/\lambda(1 + \lambda)\} \{(1 + \lambda) \log d - \log \{1 + (1 + \lambda)(d - 1)\}_+\}$$

Note  $A_{-1}(F; G) = IR_{-1}(F; G)$ ,

$$A_0(d) = 2 \{ \log d - 1 + (1/d) \} = (1/d)B_0(d).$$

### 3. Comparison Density Functions

Information divergence  $I(F; G)$  is a concept that works for both multivariate and univariate distributions. This section shows that the univariate case is distinguished by the fact that we are able to relate  $I(F; G)$  to the concept of comparison density  $d(u; F, G)$ ,

Quantile domain concepts introduced in Parzen (1979) play a central role;  $Q(u) = F^{-1}(u)$  is the quantile function. When  $F$  is continuous, we define the density quantile function  $fQ(u) = f(Q(u))$ , score function  $J(u) = -(fQ(u))'$ , and quantile density function

$$q(u) = 1/fQ(u) = Q'(u).$$

When  $F$  is discrete, we define  $fQ(u) = p_F(Q(u))$ ,  $q(u) = 1/fQ(u)$ .

The comparison density  $d(u; F, G)$  is defined as follows: when  $F$  and  $G$  are both continuous,

$$d(u; F, G) = g(F^{-1}(u))/f(F^{-1}(u));$$

when  $F$  and  $G$  are both discrete

$$d(u; F, G) = p_G(F^{-1}(u))/p_F(F^{-1}(u)).$$

In the continuous case  $d(u; F, G)$  is the derivative of

$$D(u; F, G) = G(F^{-1}(u));$$

in the discrete case we define the comparison distribution function

$$D(u; F, G) = \int_0^u d(t; F, G) dt.$$

Let  $F$  denote the true distribution function of a continuous random variable  $Y$ . To test the goodness of fit hypothesis  $H_0 : F = G$ , one transforms to  $W = G(Y)$  whose distribution

function is  $F(G^{-1}(u))$  and whose quantile function is  $G(F^{-1}(u))$ . The comparison density  $d(u; F, G)$  and  $d(u; G, F)$  are respectively the quantile density and the probability density of  $W$ .

For a density  $d(u)$ ,  $0 < u < 1$ , Renyi information (of index  $\lambda$ ), denoted  $IR_\lambda(d)$ , is non-negative and measures the divergence of  $d(u)$  from uniform density  $d_0(u) = 1$ ,  $0 < u < 1$ . It is defined:

$$IR_0(d) = 2 \int_0^1 \{d(u) \log d(u)\} du = 2 \int_0^1 \{d(u) \log d(u) - d(u) + 1\} du$$

$$IR_{-1}(d) = -2 \int_0^1 \{\log d(u)\} du = -2 \int_0^1 \{\log d(u) - d(u) + 1\} du$$

for  $\lambda \neq 0$  or  $-1$

$$IR_\lambda(d) = \{2/\lambda(1+\lambda)\} \log \int_0^1 \{d(u)\}^{1+\lambda} du$$

$$= \{2/\lambda(1+\lambda)\} \log \int_0^1 \left( \{d(u)\}^{1+\lambda} - (1+\lambda) \{d(u) - 1\} \right) du.$$

To relate comparison density to information divergence we use the concept of Renyi information  $IR_\lambda$  which yields the important identity (and interpretation of  $I(F; G)$ !)

$$I(F; G) = (-2) \int_0^1 \log d(u; F, G) du$$

$$= IR_{-1}(d(u; F, G)) = IR_0(d(u; G, F)).$$

For a density  $d(u)$ ,  $0 < u < 1$ , define

$$C_\lambda(d) = \int_0^1 B_\lambda(d(u)) du.$$

The comparison density again unifies the continuous and discrete cases. One can show that for univariate  $F$  and  $G$

$$C_\lambda(F, G) = C_\lambda(d(u; F, G))$$

**4. Approximation of positive functions (density functions) by minimum information divergence (maximum entropy)**

This section discusses how approximation theory provides models for comparison density functions. To a density  $d(u)$ ,  $0 < u < 1$ , approximating functions are defined by constraining (specifying) the inner product between  $d(u)$  and a specified function  $J(u)$ , called a score function. We often assume that the integral over  $(0,1)$  of  $J(u)$  is zero, and the integral of  $J^2(u)$  is finite. A score function  $J(u)$ ,  $0 < u < 1$ , is always defined to have the property that its inner product with  $d(u)$ , denoted

$$[J, d] = [J(u), d(u)] = \int_0^1 J(u) d(u) du,$$

is finite. The inner product is called a *component* or *linear detector*; its value is a measure of the difference between  $d(u)$  and 1.

The question of which distributions to choose as  $F$  and  $G$  is often resolved by the following formula which evaluates the inner product between  $J(u)$  and  $d(u; F, G)$  as a moment with respect to  $G$  if  $J(u) = \varphi(F^{-1}(u))$ :

$$[\varphi(F^{-1}(u)), d(u; F, G)] = \int_{-\infty}^{\infty} \varphi(y) dG(y) = E_G[\varphi(Y)]$$

Often  $G$  is a raw sample distribution and  $F$  is a smooth distribution which is a model for  $G$  according to the hypothesis being tested.

We propose that non-parametric statistical inference and density estimation can be based on the same criterion functions used for parametric inference if one uses the minimum Renyi information approach to density estimation (which extends the maximum entropy approach); form functions  $d_{\lambda, m}^{\wedge}(u)$  which minimize  $IR_{\lambda}(d^{\wedge}(u))$  among all functions  $d^{\wedge}(u)$  satisfying the constraints

$$[J_k, d^{\wedge}] = [J_k, d] \text{ for } k = 1, \dots, m$$

where  $J_k(u)$  are specified score functions. One expects  $d_{\lambda, m}^{\wedge}(u)$  to converge to  $d(u)$  as  $m$  tends to  $\infty$ , and  $IR_{\lambda}(d_{\lambda, m}^{\wedge})$  to non-decreasingly converge to  $IR_{\lambda}(d)$ .

The case  $\lambda = 1$  provides approximations in  $L_2$  norm which are based on a sequence  $J_k(u)$ ,  $k = 1, 2, \dots$ , which is a complete orthonormal set of functions. If  $d(u)$ ,  $0 < u < 1$ ,

is square integrable (equivalently,  $IR_1(d)$  is finite) one can represent  $d(u)$  as the limit of

$$d_m(u) = 1 + \sum_{k=1}^m [J_k, d] J_k(u), m = 1, 2, \dots$$

When  $\varphi_k(y)$ ,  $k = 1, 2, \dots$ , is complete orthonormal set for  $L_2(F)$ , a density  $g(y)$  can be approximated by

$$g_m(y) = f(y) \left\{ 1 + \sum_{k=1}^m E_G [\varphi_k(Y)] \varphi_k(y) \right\}$$

We call  $d_m(u)$  a truncated orthogonal function (generalized Fourier) series.

An important general method of density approximation, called a weighted orthogonal function approximation, is to use suitable weights  $w_k$  to form approximations

$$d^*(u) = 1 + \sum_{k=1}^{\infty} w_k [J_k, d] J_k(u).$$

to  $d(u)$ . Often  $w_k$  depends on a "truncation point"  $m$ , and  $w_k \rightarrow 1$  as  $m \rightarrow \infty$ .

*Quadratic Detectors.* To test  $H_0 : d(u) = 1$ ,  $0 < u < 1$ , many traditional goodness of fit test statistics (such as Cramer-von Mises and Anderson-Darling) can be expressed as quadratic detectors

$$\sum_{k=1}^{\infty} \{w_k [J_k, d]\}^2 = \int_0^1 \{d^*(u) - 1\}^2 du = C_1(d^*) = -1 + \exp IR_1(d^*).$$

We propose that these nonadaptive test statistics should be expressed as information measures and compared with minimum Renyi information detectors  $IR_{\lambda}(d_{\lambda, m}^{\wedge})$ ; in this way information can provide unification of statistical methods.

Maximum entropy approximators correspond to  $\lambda = 0$ ;  $d_{0, m}^{\wedge}(u)$  satisfies an exponential model (whose parameters are denoted  $\theta_1, \dots, \theta_m$ )

$$\log d_{0, m}^{\wedge}(u) = \sum_{k=1}^m \theta_k J_k(u) - \Psi(\theta_1, \dots, \theta_m)$$

where  $\Psi$  is the integrating factor that guarantees that  $d_{0, m}^{\wedge}(u)$ ,  $0 < u < 1$ , integrates to 1:

$$\Psi(\theta_1, \dots, \theta_m) = \log \int \exp \left\{ \sum_{j=1}^m \theta_j J_j(u) \right\} du$$

The approximating functions formed in practice are not computed from the true components  $[J_k, d]$  but from raw estimators  $[J_k, \tilde{d}]$  for suitable raw estimators  $\tilde{d}(u)$ . The approximating functions are interpreted as estimators of a true density. Methods proposed for unification and generalization of statistical methods use minimum Renyi information estimation techniques. Different applications of these methods differ mainly in how they define the raw density  $\tilde{d}(u)$  which is the starting point of the data analysis.

## 5. Equivalence and Orthogonality of Normal Time Series

This section formulates in terms of Renyi information some classic results of the theory of time series that should be part of the education of Ph.D.'s in statistics.

To apply Neyman Pearson statistical inference to a time series  $\{Y(t), t \in T\}$  with abstract index set  $T$ , we must first define the probability density functional, denoted  $p(Y(\cdot); \theta)$ . We assume a family of probability models (for the time series) parametrized by a possibly infinite dimensional parameter  $\theta$ .

The model " $Y(\cdot)$  is normal with known covariance kernel  $K(s, t) = \text{cov}\{Y(s), Y(t)\}$  and unknown mean value function  $m(t) = E[Y(t)]$ " is equivalent to a probability measure  $P_m$  on the function space  $R_T$  of all functions on  $T$ . We define  $p(Y(\cdot); m)$  to be the Radon-Nikodym derivative of  $P_m$  with respect to  $P_0$ , the probability measure corresponding to  $m(t) = 0$  for all  $t$ .

*Theorem* (Parzen (1958)): In order that  $P_m$  be absolutely continuous with respect to  $P_0$  it is necessary and sufficient that  $m$  is in  $H(K)$ , the reproducing kernel Hilbert space of functions on  $T$  with reproducing kernel  $K$  and inner product between functions  $f$  and  $g$  in  $H(K)$  denoted  $\langle f, g \rangle_{H(K)}$  and satisfying

$$\langle f, K(\cdot, t) \rangle_{H(K)} = f(t)$$

for every  $f$  in  $H(K)$  and  $t$  in  $T$ .

The probability measures  $P_m$  and  $P_0$  of normal time series are either equivalent or orthogonal; when they have the same covariance kernel  $K$  and mean value functions dif-

fering by  $m$  they are orthogonal if and only if  $\|m\|_{H(K)} = \infty$ , and they are equivalent if and only if  $\|m\|_{H(K)}$  is finite [which is equivalent to  $m$  is a member of  $H(K)$ ] and

$$\log p(Y(.); m) = \langle Y, m \rangle_{\sim H(K)} - .5 \|m\|_{H(K)}^2$$

The random variable  $\langle Y, m \rangle_{\sim H(K)}$  is a "congruence inner product" (Parzen (1970)); it is the linear combination of  $\{Y(t), t \text{ in } T\}$  corresponding to  $m$  under the congruence which maps  $K(\cdot, t)$  into  $Y(t)$ .

To compute the Renyi information of index  $\lambda$  let  $\sigma = \|m\|_{H(K)}$ ,  $\langle Y, m \rangle_{\sim H(K)} = \sigma Z$ , where  $Z$  denotes a Normal (0,1) random variable. Then

$$IR_\lambda(P_0, P_m) = \{2/\lambda(1 + \lambda)\} \log \text{Int},$$

where

$$\begin{aligned} \text{Int} &= \int_{R_T} \{p(Y(.); m)\}^{1+\lambda} dP_0 \\ &= E[\exp\{(1 + \lambda)(\sigma Z - .5\sigma^2)\}] \\ &= \exp\{.5(1 + \lambda)^2\sigma^2 - .5(1 + \lambda)\sigma^2\} \\ &= \exp\{.5\lambda(1 + \lambda)\sigma^2\} \end{aligned}$$

**Theorem:** Renyi information of two common covariance normal time series:

$$IR_\lambda(P_0; P_m) = \sigma^2 = \|m\|_{H(K)}^2$$

This beautiful formula illustrates that our definition of Renyi information has been adjusted to be equivalent to a chi-squared statistic.

Method of proof uses limits of information numbers: To prove results about equivalence and orthogonality one studies the limit of information measures of finite dimensional restrictions  $P_m^{(n)}$  representing probability measures of  $Y(t)$ ,  $t$  in a finite set  $T^{(n)}$  of points in  $T$  converging monotonely to a set  $T^{(\infty)}$  dense in  $T$ . The norm of the restriction of  $m$  to  $T^{(n)}$ , denoted  $\|m\|_{H(K, T^{(n)})}$ , may be shown to converge to  $\|m\|_{H(K)}$ . For a time



series  $\{Y(t), t \text{ in } T\}$  with abstract index set  $T$ , the finite dimensional Renyi information is denoted

$$IR_{\lambda}^{(n)} = IR_{\lambda} (P_0^{(n)}; P_m^{(n)})$$

*Theorem.* General martingale theory can be used to show that

- (1)  $P_0$  and  $P_m$  are orthogonal if and only if  $IR_{-.5}^{(n)}$  converges to  $\infty$  as  $n$  increases,
- (2)  $P_m$  is absolutely continuous with respect to  $P_0$  if  $IR_1^{(n)}$  has a finite limit as  $n$  increases.

*Proposition:* Renyi information divergence of two zero mean univariate normal distributions. Let  $P_j$  be the distribution on the real line corresponding to Normal  $(0, K_j)$  with variance  $K_j$ . Let  $p(y)$  denote the probability density of  $P_1$  with respect to  $P_2$ . Let  $\kappa = \frac{K_2}{K_1}$ . Then

$$p(y) = \kappa^{.5} \exp \left\{ -.5(\kappa - 1) \frac{y^2}{K_2} \right\}$$

$$IR_{-1}(P_2; P_1) = \kappa - 1 - \log \kappa,$$

$$IR_{\lambda}(P_2; P_1) = (1/\lambda) \{ \log \kappa - (1 + \lambda)^{-1} \log \{ 1 + (1 + \lambda)(\kappa - 1) \} \}_+$$

$$C_{\lambda}(P_2; P_1) = \{ 2/\lambda(1 + \lambda) \} \kappa^{.5(1+\lambda)} \{ 1 + (1 + \lambda)(\kappa - 1) \}_+^{-.5}$$

Asymptotic information can be computed from the fact (compare Hannan (1970), p. 429)

$$\lim_{n \rightarrow \infty} \|m\|_{H(K, T^{(n)})}^2 / \sum_{t=1}^n m^2(t) = \int_0^1 \frac{1}{R(0)f(\omega)} dM(\omega)$$

where  $M(\omega)$  is the asymptotic spectral distribution function of  $m(\cdot)$  assuming that  $m(\cdot)$  obeys Grenander's conditions, or is "persistently exciting" in the language of control engineers (Bohlin (1971)):

$$\rho_m(v) = \int_0^1 e^{2\pi i v \omega} dM(\omega),$$

where  $\rho_m(v)$  is limit of sample autocorrelations of  $m(\cdot)$ .

## 6. Asymptotic Information of Stationary Normal Time Series

This section discusses unification of information measures of stationary normal time series and information measures of non-negative functions which are spectral density functions.

When a time series  $\{Y(t), t = 1, 2, \dots\}$  is modeled by probability measures which are orthogonal over the infinite sequence but equivalent for any finite sample, we define asymptotic information divergence (or rate of information divergence)

$$\text{AsymIR}_\lambda(P_2; P_1) = \lim_{n \rightarrow \infty} (1/n) \text{IR}_\lambda(P_2^{(n)}; P_1^{(n)}).$$

Let  $Y(\cdot)$  be zero mean normal stationary with covariance function

$$R(v) = E[Y(t)Y(t-v)].$$

The correlation function is defined

$$\rho(v) = R(v)/R(0).$$

We prefer to analyze the time series after first subtracting sample mean and dividing by its sample standard deviation; its covariance function asymptotically equals  $\rho(v)$ .

An important classification of time series is by memory type: no memory, short memory, long memory according as  $I_\infty = 0$ ,  $0 < I_\infty < \infty$ ,  $I_\infty = \infty$  where

$$I_\infty = I(Y|Y_{-1}, Y_{-2}, \dots) = E_{Y_{-1}, Y_{-2}, \dots} I(f_{Y|Y_{-1}, Y_{-2}, \dots}; f_Y)$$

is the information about  $Y(t)$  in  $Y(t-1), Y(t-2), \dots$ , its infinite past (see Parzen (1981), (1983)).

Assume that  $Y(\cdot)$  is short memory and satisfies

$$\sum_{v=-\infty}^{\infty} |R(v)| \text{ finite.}$$

The spectral density function  $f(\omega)$ ,  $0 \leq \omega < 1$ , is defined as the Fourier transform of the correlation function:

$$f(\omega) = \sum_{v=-\infty}^{\infty} \exp(-2\pi i v \omega) \rho(v)$$

We call a time series *bounded* memory if the spectral density is bounded above and below:

$$0 < c_1 \leq f(\omega) \leq c_2 < \infty.$$

Let  $P_f$  denote the probability measure on the space of infinite sequences  $R_\infty$  corresponding to a normal zero mean stationary time series with spectral density function  $f(\omega)$ .

A result of Pinsker [(1964), p. 196] can be interpreted as providing a formula for asymptotic information divergence between two zero mean stationary time series with respective rational spectral density functions  $f(\omega)$  and  $g(\omega)$ . Write  $\text{AsymIR}_\lambda(f, g)$  for  $\text{AsymIR}_\lambda(P_f; P_g)$ . Adapting Pinsker (1964) one can prove that

$$\text{AsymIR}_{-1}(f, g) = \int_0^1 \{(f(\omega)/g(\omega)) - 1 - \log(f(\omega)/g(\omega))\} d\omega$$

The definition of Renyi information can be extended to non-negative functions  $d(u)$  which do not necessarily integrate to 1. Because spectral densities are even functions we take the integral to be over  $0 \leq \omega < .5$ . One obtains the following important theorem.

*Theorem:* Unification of information measures of Pinsker (1964) and Itakura-Saito (1970).

$$\text{AsymIR}_{-1}(f, g) = \text{IR}_{-1}(f(\omega)/g(\omega))_{0,.5}$$

The validity of this information measure can be extended to non-normal asymptotically stationary time series (Ephraim, Lev-Ari, Gray (1988)).

One can heuristically motivate Pinsker's information theoretic justification of the Itakura-Saito distortion measure by the formula (at the end of section 5) for the information divergence between two univariate normal distributions with zero means and different variances. Motivated by this formula we propose a formula for bounded memory time series (whose proof is given by Kazakos and Kazakos (1980)) which motivates a *new* distortion measure:  $\text{AsymIR}_\lambda(f, g) =$

$$(1/\lambda) \int_0^1 \{\log(f(\omega)/g(\omega)) - (1/(1+\lambda)) \log \{1 + (1+\lambda)((f(\omega)/g(\omega)) - 1)_+\} d\omega$$

The properties of the integrand are the same as those of  $B_\lambda(d)$ .

Kazakos and Kazakos (1980) also give formulas for asymptotic information of multiple stationary time series.

## 7. Estimation of Finite Parameter Spectral Densities

This section formulates in terms of Renyi information the classic asymptotic maximum likelihood Whittle theory of spectral estimation.

For a random sample of a random variable with unknown probability density  $f$ , maximum likelihood estimators  $\hat{\theta}$  of the parameters of a finite parameter model  $f_\theta$  of the probability density  $f$  can be shown to be equivalent to minimizing

$$IR_{-1}(f^{\sim}, f_\theta)$$

where  $f^{\sim}$  is a raw estimator of  $f$  (initially, a symbolic sample probability density formed from the sample distribution function  $F^{\sim}$ ). A similar result, called Whittle's estimator (Whittle (1953)), holds for estimation of spectral densities of a bounded memory zero mean stationary time series for which one assumes a finite parametric model  $f_\theta(\omega)$  for the true unknown spectral density  $f(\omega)$ .

A raw fully nonparametric estimator of  $f(\omega)$  from a time series sample  $Y(t)$ ,  $t = 1, \dots, n$ , is the sample spectral density (or periodogram)

$$f^{\sim}(\omega) = \left| \sum_{t=1}^n Y(t) \exp(-2\pi i \omega t) \right|^2 \div \sum_{t=1}^n |Y(t)|^2$$

Note that  $f^{\sim}(\omega)$  is not a consistent estimator of  $f(\omega)$ ; nevertheless,

$$E[f^{\sim}(\omega)] \text{ converges to } f(\omega),$$

a fact which can be taken as the definition of  $f(\omega)$ .

An estimator  $\hat{\theta}$  which is asymptotically equivalent to maximum likelihood estimator is obtained by minimizing  $\text{AsymIR}_{-1}(f^{\sim}; f_\theta) = IR_{-1}(f^{\sim}, f_\theta)_{0.5} =$

$$\int_0^1 \{ (f^{\sim}(\omega)/f_\theta(\omega)) - 1 - \log(f^{\sim}(\omega)/f_\theta(\omega)) \} d\omega$$

which can be interpreted as choosing  $\theta$  to make  $\frac{f^{\sim}(\omega)}{f_{\theta}(\omega)}$  as flat or constant as possible.

We usually use the representation

$$f_{\theta}(\omega) = \sigma^2 / \gamma_{\theta}(\omega)$$

where  $\gamma_{\theta}(\omega)$  is the square modulus of the transfer function of the whitening filter represented by the spectral density model  $f_{\theta}$ , constructed so that

$$\int_0^1 \log f_{\theta}(\omega) d\omega = \log \sigma^2.$$

Minimizing  $\text{AsymIR}_{-1}(f^{\sim}, f_{\theta})$  is equivalent to minimizing

$$(1/\sigma^2) \int_0^1 \{f^{\sim}(\omega) \gamma_{\theta}(\omega)\} d\omega + \log(\sigma^2)$$

which is equivalent to minimizing over  $\theta$

$$\sigma_{\theta}^2 = \int_0^1 \gamma_{\theta}(\omega) f^{\sim}(\omega) d\omega$$

and setting

$$\sigma^{\wedge 2} = \int_0^1 \gamma_{\theta^{\wedge}}(\omega) f^{\sim}(\omega) d\omega = \sigma_{\theta^{\wedge}}^2$$

The information divergence between the data and the fitted model is given

$$\text{IR}_{-1}(f^{\sim}, f_{\theta^{\wedge}}) = \log \sigma_{\theta^{\wedge}}^2 - \log \sigma^{\wedge 2} = I_{\infty}^{\sim} - I_{\infty}^{\wedge}$$

defining  $-I_{\infty}^{\wedge} = \log \sigma^{\wedge 2}$ ,

$$-I_{\infty}^{\sim} = \log \sigma^{\wedge 2} = \int_0^1 \log f^{\sim}(\omega) d\omega$$

This criterion (however, corrected for bias in  $I_{\infty}^{\sim}$ ) arises from information approaches to model identification (Parzen (1983)). A model fitting criterion (but not a parameter estimation criterion) is provided by the information increment

$$\begin{aligned} & I(Y | \text{all past } Y; Y \text{ values in model } \theta) \\ &= \int_0^1 -\log \{f^{\sim}(\omega) / f_{\theta^{\wedge}}(\omega)\} d\omega = \text{IR}_{-1}(f^{\sim} / f_{\theta^{\wedge}})_{0, -.5} \end{aligned}$$

One can regard it as a measure of the distance of the whitening spectral density

$$f^*(\omega) = \tilde{f}(\omega) / f_{\theta^*}(\omega)$$

from a constant function; note that  $f^*(\omega)$  is constructed to integrate to 1. When one accepts that the optimal smoother of  $f^*(\omega)$  is a constant, a "parameter-free" non-parametric estimator of the spectral density  $f(\omega)$  by a smoother of  $\tilde{f}(\omega)$  is given by the parametric estimator  $f_{\theta^*}$ . By "parameter-free" we mean that we are free to choose the parameters to make the data (raw estimator) shape up to a smooth estimator. The parameters are not regarded as having any significance or interpretation; they are merely coefficients of a representation of  $f(\omega)$ .

Portmanteau statistics to test goodness of fit of a model to the time series use sums of squares of correlations of residuals; an analogous statistic is

$$IR_1(\tilde{f}/f_{\theta^*})_{0,.5} = \log \int_0^{.5} \{\tilde{f}(\omega) / f_{\theta^*}(\omega)\}^2 d\omega$$

Goodness of fit of the model to the data (as measured by how close  $f^*(\omega)$  is to the spectral density of white noise) is the ultimate model identification criterion to decide between competing parametric models.

## 8. Minimum information estimation of spectral densities

This section provides a perspective on maximum entropy spectral estimation from the point of view of minimum Renyi information approximation.

The maximum entropy approach to the problem of spectral estimation of a stationary time series was originated by Burg (1967). It derives a parametric model  $f_{\theta}$  for the true  $f$  by imposing constraints on linear functionals of  $f$  of the form: for  $v = 0, 1, \dots, m$

$$\rho(v) = \langle \exp(2\pi i v \omega), f(\omega) \rangle_{L_2(0,1)} = \tilde{\rho}(v)$$

where  $\tilde{\rho}(v)$  are estimators (from a sample) of autocorrelations. Note that the remarkable properties of Burg's estimators derive from the fact that he first estimates in a novel

way the partial correlations and should not be interpreted as proof of the superiority of maximum entropy philosophy.

Let  $\hat{f}_m$  denote the function (among all functions  $f$  satisfying these constraints) which minimizes the neg-entropy (of order -1)

$$IR_{-1}(f)_{0.5} = \int_0^1 -\{\log f(\omega)\} d\omega$$

The solution  $\hat{f}_m$  has the following parametric form:

$$\{1/\hat{f}_m\} \text{ is linear combination of } \exp(2\pi i v \omega), v = -m, \dots, m$$

The non-negativity of  $\hat{f}$  then guarantees that  $\hat{f}$  is an autoregressive spectral density:

$$\hat{f}_m(\omega) = \sigma_m^2 \left| \sum_{j=0}^m a(j) \exp(2\pi i j \omega) \right|^{-2}$$

A negative opinion about applying autoregressive spectral estimates is expressed by Diggle (1990), p. 112: "A final method, which we mention only briefly, is to fit an  $AR(p)$  process to the data  $\{y_t\}$  and to use the fitted autoregressive spectrum as the estimate of  $f(\omega)$ . The motivation for this is threefold: fitting an autoregressive process is computationally easy, autoregressive spectra can assume a wide variety of shapes, and automatic criteria are available for choosing the value of  $p$ . Nevertheless, the method seems to fit uneasily into a discussion of what is essentially a non-parametric estimation problem. It is analogous to the use of polynomial regression for data smoothing, and is open to the same basic objection, namely that it imposes global assumptions which can lead to artefacts in the estimated spectrum."

If one's criterion is to minimize

$$IR_0(f)_{0.5} = \int_0^1 \{f(\omega) \log f(\omega)\} d\omega,$$

the neg-entropy of order 0, the solution  $\hat{f}$  obeys an exponential model (Bloomfield (1973)):

$$\{\log \hat{f}\} \text{ is linear combination of } \exp(2\pi i v \omega), v = -m, \dots, m$$

These optimization problems are related to the problem: subject to the constraints, with specified score functions  $J_k(\omega)$ ,

$$\langle f, J_k \rangle = \text{specified constant for } k = 1, \dots, m$$

find a density  $f(\omega)$ ,  $0 \leq \omega < 1$ , minimizing the  $L_p$  norm, with  $p = 1 + \lambda$ ,

$$\int_0^1 \{f(\omega)\}^{1+\lambda} d\omega$$

Theorems about this problem are given by Chui, Deutsch, and Ward (1990).

*Power correlations, inverse and cepstral correlations:* Parametric models for the spectral density  $f$  can be obtained from various maximum entropy criteria. To check which model is parsimonious, one requires goodness of fit procedures which check the significant difference from zero of the Fourier transforms of various functions of  $f$  such as  $1/f$ ,  $\log f$ , or  $f^\lambda$ .

Let  $e_v(\omega) = \exp(2\pi i j \omega)$ , and interpret inner products as  $L_2(0,1)$ . Define:  
inverse correlations

$$\rho^{(-1)}(v) = \langle e_v, 1/f \rangle$$

cepstral correlations

$$\rho^{(0)}(v) = \langle e_v, \log f \rangle$$

ordinary correlations

$$\rho^{(1)}(v) = \langle e_v, f \rangle$$

power correlations of power  $\lambda$

$$\rho^{(\lambda)}(v) = \langle e_v, f^\lambda \rangle$$

Identification of a parametric model for  $f$  should include routine estimation and interpretation of these various correlations.

In general if one expands  $f^\lambda(\omega)$  as a linear combination of orthogonal functions  $J_k(u)$ ,  $0 \leq \omega < 1$ , one forms the transform (called power orthogonal series coefficients)

$$\rho^{(J,\lambda)}(k) = \langle J_k, f^\lambda \rangle$$



An open research problem is identification of appropriate orthogonal functions  $J_k(u)$ ,  $0 \leq \omega < 1$ .

## 9. Tail Classification of Probability Laws and Spectral Densities.

This section discusses models for spectral density functions which are based on their analogy with quantile density functions.

From extreme value theory, statisticians have long realized that it is useful to classify distributions according to their tail behavior (behavior of  $F(x)$  as  $x$  tends to  $\pm\infty$ ). It is usual to distinguish three main types of distributions, called (1) limited, (2) exponential, and (3) algebraic. Parzen (1979) proposes that this classification be expressed in terms of the density quantile function  $fQ(u)$ ; we call the types short, medium, and long tail.

A reasonable assumption about the distributions that occur in practice is that their density-quantile functions are *regularly varying* in the sense that there exist tail exponents  $\alpha_0$  and  $\alpha_1$  such that, as  $u \rightarrow 0$ ,

$$fQ(u) = u^{\alpha_0} L_0(u), \quad fQ(1-u) = u^{\alpha_1} L_1(u)$$

where  $L_j(u)$  for  $j = 0, 1$  is a slowly varying function.

A function  $L(u)$ ,  $0 < u < 1$  is usually defined to be *slowly varying* as  $u \rightarrow 0$  if, for every  $y$  in  $0 < y < 1$ ,  $L(yu)/L(u) \rightarrow 1$  or  $\log L(yu) - \log L(u) \rightarrow 0$ . For estimation of tail exponents we will require further that, as  $u \rightarrow 0$ ,

$$\int_0^1 \{\log L(uy) - \log L(u)\} dy \rightarrow 0$$

which we call *integrally slowly varying*. An example of a slowly varying function is  $L(u) = \{\log u^{-1}\}^\beta$ .

*Classification of tail behavior of probability laws.* A probability law has a left tail type and a right tail type depending on the value of  $\alpha_0$  and  $\alpha_1$ . If  $\alpha$  is the tail exponent, we

define:

$\alpha < 0$	super short tail
$0 \leq \alpha < 1$	short tail
$\alpha = 1$	medium tail
$\alpha > 1$	long tail

Medium tailed distributions are further classified by the value of  $J^* = \lim J(u)$ :

$\alpha = 1, J^* = 0$	medium long tail
$\alpha = 1, 0 < J^* < \infty$	medium-medium tail
$\alpha = 1, J^* = \infty$	medium-short tail

One immediate insight into the meaning of tail behavior is provided by the hazard function  $h(x) = f(x) \div \{1 - F(x)\}$  with hazard quantile function  $hQ(u) = fQ(u) \div 1 - u$ . The convergence behavior of  $h(x)$  as  $x \rightarrow \infty$  is the same as that of  $hQ(u)$  as  $u \rightarrow 1$ . From the definitions one sees that  $h^* = \lim_{x \rightarrow \infty} h(x)$  satisfies

$h^* = \infty$ (increasing hazard rate)	Short or medium-short tail
$0 < h^* < \infty$ (constant hazard rate)	Medium-medium tail
$h^* = 0$ (decreasing hazard rate)	Long or medium-long tail

*Formulas for computing tail exponents.* The representation of  $fQ(u)$  suggests a formula for computation of tail exponents  $\alpha_0$  and  $\alpha_1$  (which may be adapted to provide estimators from data):

$$-\alpha_0 = \lim_{u \rightarrow 0} \int_0^1 \{\log fQ(uy) - \log fQ(u)\} dy$$

$$\alpha_1 = \lim_{u \rightarrow 0} \int_0^1 \{\log fQ(1 - yu) - \log fQ(1 - u)\} dy$$

*Memory classification of spectral densities:* Spectral densities with no poles or zeroes represent time series with bounded memory. We regard spectral density functions as

analogous to quantile density functions. A model for a spectral density with a pole or zero at zero frequency (a similar representation holds for an arbitrary frequency  $\omega_0$ ) is (Parzen (1986))

$$f(\omega) = \omega^{-\delta} L(\omega)$$

where  $L$  is a slowly varying function at  $\omega = 0$  and  $L(0) > 0$ . An important role is played by  $f(1/n) = n^\delta L(1/n)$ .

The spectral density is integrable if  $\delta < 1$ , which is the condition for stationarity. The spectral density of a non-stationary time series needs careful definition. The case  $\delta = 1$  is of particular interest; it corresponds to "1/f" noise. The case  $\delta > 1$  could be called "fractal noise". A time series whose first difference is stationary has  $\delta = 2$ . Heuristically,  $\delta$  is interpreted for a zero mean time series  $Y(\cdot)$  by

$$E \left[ \frac{1}{n} \left| \sum_{t=1}^n Y(t) \right|^2 \right] \text{ grows as } n^\delta L(1/n).$$

The index  $\delta$  associated with frequency  $\omega$  is interpreted:

$$E \left[ \frac{1}{n} \left| \sum_{t=1}^n \exp(2\pi i \omega t) Y(t) \right|^2 \right] \text{ grows as } n^\delta L(1/n),$$

and, when  $\delta = 0$ , converges to  $R(0)f(\omega)$  if it is finite.

This approach provides definitions of spectral density for asymptotically stationary time series (Parzen (1962)).

Note that a finite dynamic range (bounded memory) spectral density has  $\delta = 0$ , but  $\delta = 0$  does not imply finite dynamic range since  $f(\omega)$  can tend to  $\infty$  as  $\omega \rightarrow 0$ ; an example is  $f(\omega) \sim (\log \omega)^2$ ,  $\rho(v) \sim (\log v)/v$  as  $v \rightarrow \infty$ .

A traditional parametrization of stationary long memory time series is  $\delta = 2H - 1$ , where  $H$  is the Hurst index satisfying  $.5 < H < 1$ ; Hurst estimated  $H = .7$  for the Nile water level time series. The covariance function has the asymptotic representation

$$R(v) \text{ decays slowly like } v^{2H-2} = v^{\delta-1}.$$

The memory index delta plays an important theoretical role. In many time series theorems the asymptotic behavior of a statistic is expressed in terms of  $f(0)$ , the value at zero frequency of the spectral density function. These results often have analogies for long memory time series if one replaces  $f(0)$  by  $f(1/n) = n^\delta$  asymptotically; compare Samarov and Taqqu (1988).

Estimation of delta can be considered estimating a "fractal dimension", the exponent of the rate of growth of the mean of the sample spectral density. Values of delta are used to describe music and how the brain works! *U. S. News and World Report*, June 11, 1990, p. 62 writes: "Surprisingly, the same mathematical formula that characterizes the ebb and flow of music has been discovered to exist widely in nature, from the flow of the Nile to the beating of the human heart to the wobbling of the earth's axis. Remarkably, this equation is closely related to other mathematical formulas used by computer experts to generate amazingly realistic pictures of coastlines, clouds and mountain ranges and other natural scenery."

Estimating delta from data has many of the same difficulties as estimating the tail index of a probability distribution. Since delta is a property of a long memory time series, it undoubtedly can not be estimated with great accuracy from relatively short lengths of observed time series.

We would appreciate references to research about delta, especially the conjectures in this section.

#### 10. Sample Brownian Bridge Exploratory analysis of time series.

To a time series sample  $\{Y(t), t = 1, 2, \dots, n\}$  one can associate functions  $d^*(u)$ ,  $0 \leq u \leq 1$ , and

$$D^*(u) = \int_0^u d^*(t) dt$$

satisfying  $D^*(1) = 0$ . Let  $\mu_n$  and  $\sigma_n$  denote respectively the sample mean and sample

standard deviation. Define, for  $j = 1, \dots, n$ ,

$$d^{\sim}(u) = \sqrt{n} (Y(j) - \mu_n) / \sigma_n, \quad \frac{j-1}{n} < u < \frac{j}{n}$$

Note that for  $k = 1, \dots, n$

$$D^{\sim}\left(\frac{k}{n}\right) = \frac{1}{\sqrt{n}} \sum_{t=1}^k (Y(t) - \mu_n) / \sigma_n$$

We call  $D^{\sim}(\cdot)$  the *sample Brownian Bridge* of an observed time series. We propose that a time series analysis should routinely examine the graph of  $D^{\sim}(u)$ ,  $0 \leq u < 1$ ; one can show by examples that it provides graphical tools of identification of various types of long memory time series.

*Theorem:* For a stationary time series with bounded memory

$$\{D^{\sim}(u), 0 \leq u \leq 1\} \text{ converges in distribution to } \{f^{.5}(0)B(u), 0 \leq u \leq 1\},$$

where  $B(u)$ ,  $0 \leq u \leq 1$ , is a Brownian Bridge stochastic process and  $f(0)$  is the spectral density at zero frequency.

For long memory time series we would like to understand the asymptotic behavior of

$$\{f(1/n)\}^{-.5} D^{\sim}(u), \quad 0 \leq u \leq 1.$$

*Simulation and Time Series:* Note that similar processes are studied by researchers (Schruben, Iglehart) in operations research departments who study simulation methods of forming confidence intervals for  $\mu$ , the true mean of  $Y(\cdot)$ ; they standardize  $D^{\sim}(u)$  by its maximum minus its minimum.

*Quality Control and Time Series:* The process  $D^{\sim}(\cdot)$  also has applications to quality control problems of identifying departures from the null hypothesis that  $Y(\cdot)$  is white noise. Components (linear functionals) of  $d^{\sim}(\cdot)$  are related to accumulation analysis methods of Taguchi.

## REFERENCES

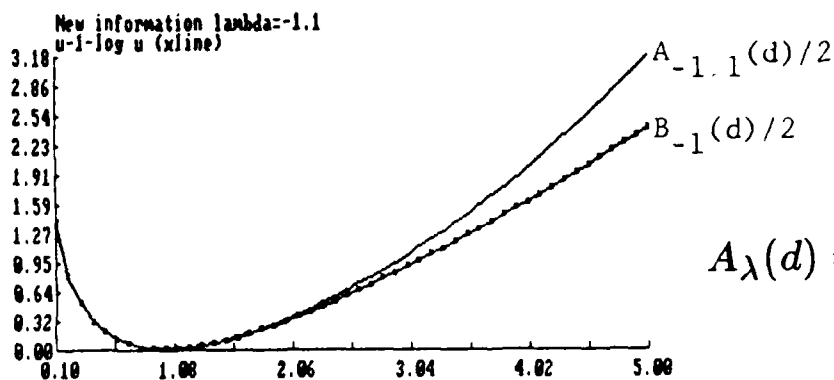
- Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. Proc. of the Second International Symposium on Information Theory, B. N. Petrov and F. Csaki, Akademiai Kiado, Budapest, pp. 267-281.
- Akaike, H. (1974). "A new look at the statistical model identification," *IEEE Trans. Autom. Contr.*, AC-19, 716-723.
- Akaike, H. (1977). "On entropy maximization principle," *Application of Statistics* (Krishnaiah ed.), North-Holland, 27-41.
- Akaike, H. (1978). "Canonical correlation analysis of time series and the use of an information criterion," *System Identification: Advances and Case Studies* (R. K. Mehra and D. G. Lainiotics, eds.), Academic Press: 27-96.
- Akaike, H. (1980). "Likelihood and the Bayes procedure," *Bayesian Statistics* (Bernardo, J. M., De Groot, M. H., Lindley, D. U. and Smith, A. F. M. eds.). University Press: Valencia, Spain.
- Akaike, H. (1983). "Information measures and model selection", *Proc. International Statistical Institute* (44th Session, Madrid) 1, 227-291.
- Akaike, H. (1983). "Statistical inference and measurement of entropy" in *Scientific Inference, Data Analysis, and Robustness*, ed. G. E. P. Box and C. J. Wu, Academic Press: New York.
- Ali, S. M., Silvey, S. D. (1966). "A general class of coefficients of divergence of one distribution from another. *J. Roy. Stat.*, 28, 131-142.
- Arimoto, S. (1971). "Informational-theoretical considerations on estimation problems *Inf. and Control*, 19, 181-194.
- Bhattacharyya, A. (1943). "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp 99-109.
- Bloomfield, P. (1973). "An exponential model for the spectrum of a scalar time series. *Biometrika*, 60, No. 2, 217-226.
- Bohlin, T. (1971). "On the problem of ambiguities in maximum likelihood identification," *Automatica*, vol. 7, pp. 199-210.
- Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control*, 2nd edition. Holden Day, San Francisco.
- Brillinger, D. R. (1981). *Time Series*. Holden Day, San Francisco.
- Brockett, P., Charnes A. and Cooper, W. W. (1980). "MDI estimation via unconstrained convex programming." *Comm. Statist. B-Simulation Comput.*, 9, pp. 223-234.
- Brockwell, P. J. and Davis, R. A. (1987). *Time Series: Theory and Methods*, Springer-Verlag, New York.
- Burg, J. P. (1967). "Maximum entropy spectral analysis," in *Proc. 37th Meeting Soc. of Exploration Geophysicists*. Oklahoma City, OK.
- Caines, Peter E. (1988). *Linear Stochastic Systems*. Wiley, New York.
- Charnes, A., Cooper, W. W. and Seiford, L. (1978). "Extremal principles and optimization qualities for Khinchin-Kullback-Leibler estimation," *Math Operationsforsch, Statist.*, vol. 9, pp. 21-29.

- Charnes, A., Cooper, W. W., and Tyssedal, J. (1983). "Khinchin-Kullback-Leibler estimation with inequality constraints," *Math. Operationsforsch. Statis. Ser. Optim.*, 14, pp. 1-4.
- Chui, C. K., Deutsch, F., Ward, J. D. (1990). "Constrained best approximation in Hilbert space," *Constructive approximation*, 67, 35-64.
- Csiszár, I. (1967). "Information-type measures of divergence of probability distributions and indirect observations", *Studia Sci. Math. Hung.* 2, 299-318.
- Csiszár, I. (1975). "I-divergence geometry of probability distributions and minimization problems." *Ann. Probab.*, vol. 3, pp. 146-168.
- David, H. T. and Kim, Geung-ho. (1979). "Pragmatic optimization of information functionals," in *Optimizing Methods in Statistics*, ed. Jagdish S. Rustagi, pp. 167-181. Academic Press: New York.
- Diggle, P. J. (1990). *Time Series: A Biostatistical Introduction*, Clarendon Press: Oxford.
- Ephraim, Y., Hanoach, L., and Gray, R. (1988). "Asymptotic minimum discrimination information measure for asymptotically weakly stationary processes," *IEEE Transactions on Information Theory*, vol. 34, No. 5, pp. 1033-1040.
- Gersch, W., et al. (1979). "Automatic classification of electroencephalograms," *Science*, vol. 205, pp. 193-195.
- Guiasu, S. (1977). *Information Theory with Application*, Great Britain, McGraw-Hill Book. Co., Inc.
- Hannan, E. J. (1970). *Multiple Time Series*. New York: Wiley.
- Itakura, F. and Saito, S. (1968). "Analysis synthesis telephony based on the maximum likelihood method," in *Proc. 6th. Int. Conf. Acoustics.*, Tokyo, Japan. pp. C17-C20.
- Itakura, F. and Saito, S. (1970). "A statistical method for estimation of speech spectral density and format frequencies," *Electron. Commun. Japan*, 53-A, 36-43.
- Jones, L. K. (1989). "Approximation theoretic derivation of logarithmic entropy principles for inverse problems and unique extension of the maximum entropy method to incorporate prior knowledge," *SIAM J. Appl. Math.*, vol. 49, pp. 650-661.
- Jones, L. K. and Byrne, C. L. (1990). "General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis," *IEEE Transactions on Information Theory*, vol. 36, no. 1, pp. 23-30.
- Jones, L. K. and Trutzer, V. (1989). "Computationally feasible high resolution minimum distance procedures which extend the maximum entropy method," *Inverse Problems*, vol. 5, pp. 749-766.
- Kallianpur, G. (1960). "On the amount of information in a sigma field", in Ingram Olkin et al, editors, *Contributions to Probability and Statistics in Honor of H. Hotelling*, Stanford University Press.
- Kazakos, D. and Papantoni-Kazakos, P. (1980). "Spectral distance measures between Gaussian processes," *IEEE Trans. Automat. Contr.*, vol. AC-25, no. 5, pp. 950-959 .
- Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.
- McClellan, J. (1982). "Multidimensional spectral estimation," *Proc. IEEE*, vol. 70, pp. 1029-1039.
- Mandelbrot, B. B. and Taqqu, M. S. (1979). "Robust R/S analysis of long run serial correlation." Proceedings of the 42nd session of the International Statistical Institute,

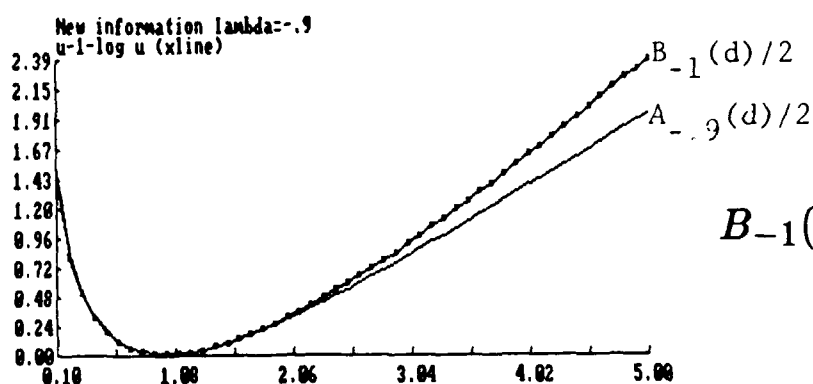
- Manila. *Bull. I.S.I.* 48 (Book 2), 69-104.
- Newton, H. J. (1988). *TIMESLAB: A Time Series Analysis Laboratory*, Wadsworth: Pacific Grove, California.
- Parzen, E. (1967). *Time Series Analysis Papers*, Holden-Day, San Francisco, California.
- Parzen, E. (1958) "On asymptotically efficient consistent estimates of the spectral density function of a stationary time series" *J. Royal Statist. Soc., Ser. B.*, 20, 303-322.
- Parzen, E. (1959) "Statistical Inference on Time Series by Hilbert Space Methods, I" Technical Report 23, January 2, 1959, Statistics Department, Stanford University. Reprinted in *Time Series Analysis Papers*.
- Parzen, E. (1961). "An approach to time series analysis," *Ann. Math. Statist.*, 32, 951-989.
- Parzen, E. (1962). "Spectral analysis of asymptotically stationary time series", *Bull. Inst. Internat. Statist.*, 39, 2, 87-103.
- Parzen, E. (1969). "Multiple time series modeling" *Multivariate Analysis - II*, edited by P. Krishnaiah, Academic Press: New York, 289-409.
- Parzen, E. (1970). "Statistical inference on time series by RKHS methods, II" *Proceedings of the 12th Biennial Seminar of the Canadian Mathematical Congress*, edited by R. Pyke, Canadian Mathematical Congress, Montreal, 1-37.
- Parzen, E. (1971). "On the equivalence among time series parameter estimation, approximation theory and control theory", *Proceedings of the Fifth Princeton Conference on Information Science*, Princeton, 1-5.
- Parzen, E. (1974). "Some Recent Advances in Time Series Modeling", *IEEE Transactions on Automatic Control*, AC-19, 723-730.
- Parzen, E. (1977). "Multiple Time Series: Determining the Order of Approximating Autoregressive Schemes", *Multivariate Analysis - IV*, Edited by P. Krishnaiah, North Holland: Amsterdam, 283-295.
- Parzen, E. (1979). "Nonparametric Statistical Data Modeling", *Journal of the American Statistical Association*, (with discussion), 74. 105-131.
- Parzen, E. (1981). "Time Series Model Identification and Prediction Variance Horizon," *Proceedings of Second Tulsa Symposium on Applied Time Series Analysis*. Academic Press: New York. 425-447.
- Parzen, E. (1982). "Maximum Entropy Interpretation of Autoregressive Spectral Densities," *Statistics and Probability Letters*, 1, 2-6.
- Parzen, E. (1983). "Time Series Model Identification by Estimating Information," *Studies in Econometrics, Time Series, and Multivariate Statistics in Honor of T. W. Anderson*, ed. S. Karlin, T. Amemiya, L. Goodman, Academic Press: New York, 279-298.
- Parzen, E. (1983). "Time Series ARMA Model Identification by Estimating Information" *Proceedings of the 15th Annual Symposium on the Interface of Computer Science and Statistics*, Amsterdam: North Holland.
- Parzen, E. (1983). "Time Series Model Identification by Estimating Information, Memory, and Quantiles," *Questo*, 7, 531-562.
- Parzen, E. (1986). "Quantile Spectral Analysis and Long Memory Time Series," *Journal of Applied Probability*, Vol. 23A, 41-55.



- Parzen, E. (1990). "Statistical Culture," University Faculty Lecture, Texas A&M University.
- Parzen, E. (1990). "Unification of statistical methods for continuous and discrete data," (Preliminary report), Technical report no. 105, Department of Statistics, Texas A&M University.
- Pinsker, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. San Francisco, CA: Holden-Day.
- Priestley, M. B. (1981). *Spectral Analysis and Time Series*, Academic Press: London.
- Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness of Fit Statistics for Discrete Multivariate Data*, Springer Verlag, New York.
- Renyi, A. (1961). "On measures of entropy and information." *Proc. 4th Berkeley Symp. Math. Statist. Probability*, 1960, 1, 547-561. University of California Press: Berkeley.
- Renyi, A. (1967). "On some basic problems of statistics from the point of view of information theory, *Proc. 5th Berkeley Symp. on Math., Stat. and Probability*, 531-543.
- Renyi, A. (1967). "Statistics and information theory", *Studia Sci. Math. Hungarica* 2, 249-256.
- Renyi, A. (1970). *Probability Theory*. North-Holland Publ. Amsterdam.
- Rosenblatt, M. (1985). *Stationary Sequences and Random Fields*, Birkhauser: Boston.
- Samarov, A., Taquq, M. S. (1988). "On the efficiency of the sample mean in long-memory noise," *Journal of Time Series Analysis*, 9, 191-200.
- Shore, J. (1981). "Minimum cross-entropy spectral analysis," *IEEE Trans Acoust. Speech, Signal Processing*, vol. ASSP-29, no. 2, pp. 230-237.
- Shore, J. and Gray, R. (1982). "Minimum cross-entropy pattern classification and cluster analysis," *IEEE Trans. Pattern. Anal. Machine Intell.*, PAMI-4, 11-17.
- Shore, J. and Johnson, R. (1980). "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inform. Theory*. IT-26, 26-37.
- Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1983). *Akaike Information Criterion Statistics*, D. Reidel: Boston.
- Smith, C. R. and Erickson, G. J. (1987). *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, D. Reidel: Boston.
- Theil, H. and Fiebig, D. C. (1984). *Exploiting Continuity: Maximum Entropy Estimation of Continuous Distributions*, Ballinger, Cambridge, MA.
- Tukey, J. W. (1984). *Collected Works on Time Series*, Vol. I: 1949-1964, Vol. II: 1965-1984. Ed. D. R. Brillinger, Wadsworth: Belmont, California.
- Whittle, P. (1953). "Estimating and information in stationary time series." *Ark. Math.* 2, 423-434.
- Whittle, P. (1953). "The analysis of multiple stationary time series," *J. Royl Statist. Soc., B.* 15:125-139.
- Woods, J. W. (1976). "Two-dimensional Markov spectral estimation," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 5.
- Zhang, J and Brockett, P. (1987). "Quadratically constrained information theoretic analysis," *SIAM J. Appl. Math.*, vol. 47, no. 4, pp 871-885.



$$A_{\lambda}(d) = \{2/\lambda(1+\lambda)\} \{(1+\lambda) \log d - \log \{1 + (1+\lambda)(d-1)\}_+\}$$



$$B_{-1}(d) = -2 \{ \log d - d + 1 \}$$

